Reading in a Foreign Language April 2008, Volume 20, No. 1
ISSN 1539-0578 pp. 43–69

Comparing the vocabulary of different graded-reading schemes

Udorn Wan-a-rom
Mahasarakham University
Thailand

Abstract

This study compared graded-reader wordlists with the General Service List (GSL; West, 1953) and investigated the words in those lists and the words actually used in graded-reader books. The wordlists from the 2 major graded-reader series, the GSL, and the words actually used in the graded readers were examined using the Range program. The comparisons showed that the lists are different from each other largely because of the different sizes of the lists and because of the words they contain and do not contain. In addition, the words actually used in the books do not stick closely to the words in the lists on which they are based, especially at Level 1. Conclusions and implications are drawn for practice in extensive reading programs.

Keywords: graded levels, graded-reading schemes, graded readers, wordlists, extensive reading

A graded-reader scheme usually has word and structure lists that are divided into levels to guide writers and editors in designing graded-reading books. The findings of Nation and Wang's (1999) research show that most graded-reader schemes set up conditions that will enhance vocabulary learning. The limited vocabulary at each level will be repeated in books of the same level. Words from earlier levels will be repeated very often at subsequent levels, and this will provide learners with more opportunities to encounter the words. These repetitions are believed to be crucial for establishing word knowledge. According to Nation and Wang, about 10 repetitions are desirable, but the more the better.

Nation and Wang (1999) also found that 84.7% of the words in the General Service List (GSL; West, 1953) appeared in the Oxford Bookworms' (OBW) lists,[1] showing that the classic list of the 2,000 GSL words is of practical use to writers of graded readers. A general-service vocabulary is essential for all learners, no matter the modes in which and purposes for which they are using English as a foreign or second language. This claim is supported by the finding that the GSL provides around 82% average coverage of various kinds of written texts (Hirsh & Nation, 1992; Hwang & Nation, 1989; Sutarsyah, Nation, & Kennedy, 1994),[2] with higher coverage for more informal text. However, learners need vocabulary sizes that will cover at least 98% of the texts they read (Hu & Nation, 2000). According to Nation (2006), for unsimplified texts, this would require a vocabulary size of approximately 7,000–9,000 word families (i.e., headwords together with their other common forms). The notion of vocabulary size has been taken as a guideline for devising a scheme for graded readers. Ideally, graded-reading schemes

would take learners step by step with 98% coverage at each step until they can read unsimplified text with the same coverage. Unfortunately, as Nation and Wang (1999) showed, most schemes of graded readers are not well designed in terms of vocabulary size.

Frequency counts of English substantially agree on the high-frequency words (Nation, 2004). Because the levels of graded readers make use of these high-frequency words, the various wordlists of graded readers are likely to be composed of substantially the same words. Designed as readable texts for second language learners, graded readers use a controlled vocabulary and structural features that are arranged in stages or levels of increasing difficulty. These stages or levels form graded-reading schemes. The primary purpose of the wordlists in these schemes is to provide guidelines for writers and editors of graded readers. Publishers usually set the different levels for graded readers according to the number of headwords, and writers can use a wide range of words in the lists, depending on the story or topic. Presumably, vocabulary is selected chiefly on the basis of frequency, but the wordlist may be modified for a particular title based on the requirements of the story. Different publishers cannot be guaranteed to make lists with the same words and with the same number of headwords at the same level.

Because of this, no systematic comparison of the levels of the various schemes has been made beyond reviews every few years, which have dealt with content, features, and the number of headwords appearing in the catalogues to compare different schemes, as in Hill's (1997, 2001) reviews of graded readers. However, these reviews did not examine the wordlists in detail in terms of the words in the wordlists and the actual words used in the books.

Although many of the graded-reader series on the world market probably depend on West's (1953) GSL as a basis for the choice of words used in the books, for commercial purposes, the publishers have produced wordlists of their own, which are likely to be confidential and unique. Various wordlists have resulted, and the words included and the number of levels vary with the grading scheme. Little is known about the similarities of the wordlists. One way to check this is to compare the wordlists of the series to determine the amount of overlap between the lists.

The purpose of this study was to examine the wordlists of graded readers in detail. This should answer the question of whether the lists from the various series are similar enough to use as a basis for setting up reading schemes for an extensive reading program or reading across series, which pertains to language learning in general and vocabulary in particular.

The study compared sets of wordlists of two major series: those of the OBW by the Oxford University Press and the Cambridge English Readers (CER) by the Cambridge University Press. It also looked at the amount of overlap between the words in the two series and the GSL words. The results of this study are discussed to answer three questions:

1. How similar are the lists?

2. How is the GSL related to the lists?

3. Do the books at specific grade levels follow the lists designed for these levels?

The Range program and manual methods were used to compare the lists.

Method

The Computer Program

The Range program is a Windows-based program developed by Paul Nation and Alex Heatley (2002) of the Victoria University of Wellington and is freely downloadable. It can be used with three distinct word lists, called *baseword* lists, on any text. The baseword lists contain word families. For example, the headword *ABLE* is grouped with its family members *abler*, *ablest*, and *ably*. Thus, the three family members are counted as the same word, *ABLE* (see Appendix A). The Range program can sort a text's vocabulary into three categories of word families from each list and a category of words outside all the three lists, making four categories altogether (see Appendix B). The program can do this either by *range* across several texts or by *frequency* within a text. It can also mark each word according to the category in which it falls. The baseword lists can be altered depending on specific requirements. The ones that come with the program are the first and second thousand words from the GSL and Averil Coxhead's (2000) Academic Word List (AWL).[3] The program has self-checking routines to ensure that a word form does not occur in more than one of the baseword lists. This program has been used with the text-based studies of Hirsh and Nation (1992), Laufer and Nation (1995), Coxhead (2000), Chung and Nation (2004), and Nation (2006).

Graded-Reader Schemes

Although Oxford University Press, Cambridge University Press, Pearson Education, and Macmillan Education are four of the largest internationally recognized publishers of graded readers on the world market, this study only used the wordlists of the two series (i.e., OBW and CER) by the first two publishers because they were willing to provide the wordlists for the graded-reader schemes. For commercial reasons, the wordlists for the Penguin and Macmillan readers are not released to the public.

Procedure

The two series both contain six levels. Because the study involved comparing words in the wordlists, the six original wordlists from each series, which are in lemmas, had to contain the same kinds of word families. To obtain good matches between the word families in the lists of the two series, a standard set of word families had to be made.

Step I: Investigation and modification of the words in the original lists. The words in the original publishers' lists of the two series are marked with parts of speech, and each word is marked with a number to indicate the level where it occurs. For example, "1 *slow* (adj.)" means the word *slow* occurs at Level 1 as an adjective. The original publishers' lists did not include numbers, days of the week, and months of the year. When the actual books were checked, these words were found to be used, and some letters of the alphabet and abbreviations were used as well. Nation and Wang (1999) also noted that such words were freely used in graded readers at all levels.

Therefore, they were added to the original publishers' wordlists. The numbers included both the cardinal and ordinal numbers, their plural forms (*threes*, *thirds*), and the abbreviations *rd*, *st*, and *th* of the ordinal numbers.

Step II: Construction of the baseword lists. In this study, a word family is defined according to the idea put forward by Bauer and Nation (1993). A word family consists of a baseword and all the derived and inflected forms that can be understood by a learner without having to learn each form separately. Bauer and Nation used frequency, productivity, and regularity as the criteria for establishing the various levels of a word family. Level 3 of the Bauer and Nation scheme was used because this includes all the inflected forms and a small group of high-frequency, regular, and productive derived forms. This level seemed most suitable for the proficiency of the learners who would be reading the graded readers.

The inflectional categories are *plural*, *third person singular present tense*, *past tense*, *past participle*, *-ing*, *comparative*, *superlative*, and *possessive*.

The derivational affixes allowed at Level 3 are *-able*, *-er*, *-ish*, *-less*, *-ly*, *-ness*, *-th*, *-y*, *non-*, and *un-*, all with restricted uses. The following examples are of families at Level 3:

ACTOR: ACTORS
CLEAR: CLEARED, CLEARS, CLEARING, CLEARINGS, CLEARER, CLEAREST, CLEARLY, CLEARNESS
BREAK: BREAKS, BROKE, BROKEN, BREAKABLE, UNBREAKABLE, UNBROKEN, BREAKING
NINETY: NINETY, NINETIETH, NINETIETHS, NINETIES

Abbreviations such as the following are located under their word families.

ROAD: RD
STREET: ST
MOUNTAIN: MT
FEBRUARY: FEB
VOLUME: VOL

The Oxford and Cambridge wordlists were modified according to the following criteria: (a) The same words in both lists must have the same family members; (b) a family member in one list cannot be a headword in another list; and (c) a compound word in both lists is treated in a similar manner, that is, a hyphen is taken out to let the basewords stand alone or the word is used without a space or a hyphen in both lists.

A major weakness of the Range program is that it deals with word forms. Thus, it was not able to distinguish words' parts of speech and meanings, namely, words that had the same written forms but different meanings; for example, *march* (n.) and *march* (v.) were recognized as the same word by the program. This problem also occurred with most words that do not change their written forms to indicate tense such as *put* and *shut.* However, the latter problem does not matter much because whether such verbs are in present or past, they do not change meaning and are members of the same families. The same problems were found in both the wordlists compared.

Step III: Comparison of the wordlists. Before the wordlists of the two series were compared, the baseword lists of the two series were carefully checked to make sure that all the words in the lists were included at the levels intended by the publisher and that they all had the same family members:

1. The baseword lists of the two series were run against the publishers' lists to check accurate matching of the headwords and the words in the publishers' lists.

2. The baseword lists of each series were then combined, and that combined list was used to make sure that the same family members were included under the same headwords. This was to check that all family members under the same headwords in the two sets of baseword lists were the same.

3. The six new baseword lists for each series were constructed from the combined lists. Reestablishing the six new baseword lists after rechecking all headwords and their family members in the combined lists avoided some errors that might happen with some headwords in either lists.

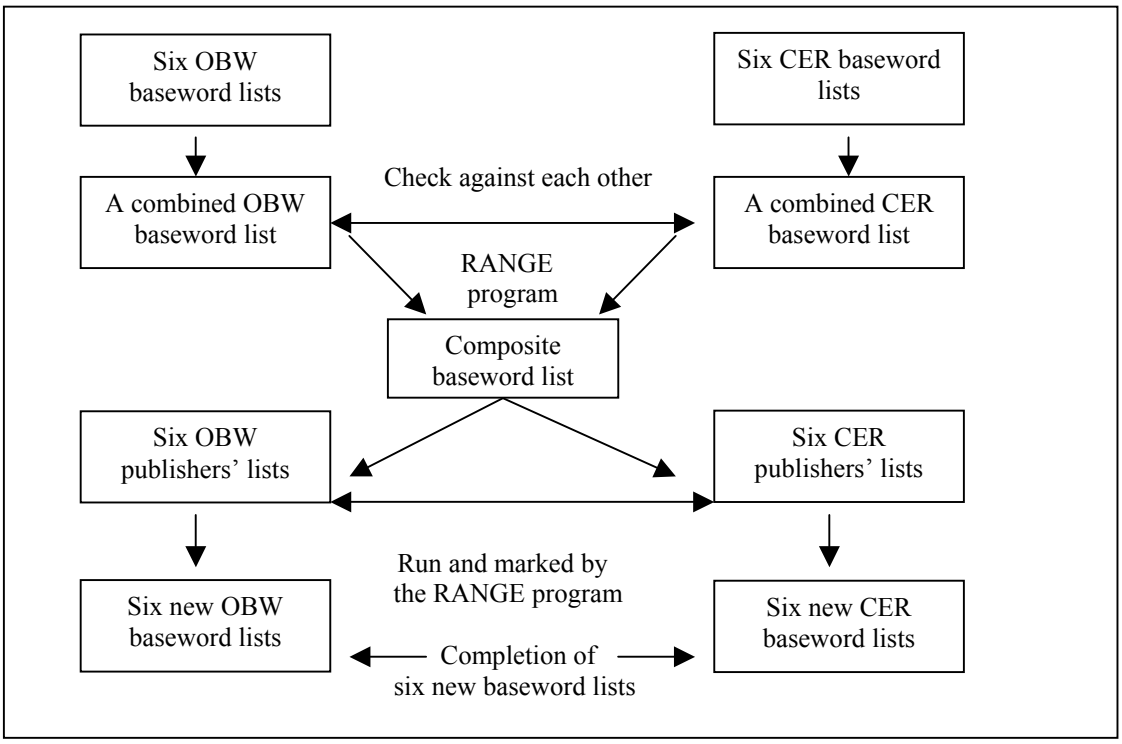


Figure 1. A flowchart of how the six comparable baseword lists for each series were constructed.

The result was two sets of six levels of baseword lists (one set each for Oxford and Cambridge) that included all the words at the right levels with the same family members for each word family (see Appendix C). This procedure is shown in Figure 1. Then, the OBW and CER baseword lists were used to compare and check words both in the original wordlists and in books of the two series.

Results

Comparing the Number of Levels and the Numbers of Words at Each Level

Three findings about the two lists are shown in Table 1: (a) the number of levels, (b) the number of word families at each level, and (c) the total number of word families used in the two series. Each series has six levels, and this makes the wordlists easier to compare. In total, the OBW list includes 2,257 word families, and the CER, 3,055. Thus, the CER list contains about 800 more word families than the OBW list. However, the numbers of new and total word families at the lower levels (1–3) are very similar between the two lists, with only small differences in the numbers of families. At Level 3, the difference is only six families. At Levels 4–6, the CER lists introduce many more families than the OBW lists, and the differences between the two lists are much larger.

Table 1. *Number of word families at each level and cumulative totals for the OBW and CER lists*

Level	New word families		Cumulative word families		Difference in new word families (OBW-CER)
	OBW	CER	OBW	CER	
1	496	477	496	477	19
2	328	320	824	797	27
3	306	339	1,130	1,136	-6
4	273	502	1,403	1,638	-235
5	423	670	1,826	2,308	-482
6	431	747	2,257	3,055	-798

Although the numbers of word families are very similar in Levels 1–3, the families at these lower levels may not in fact be the same in both lists. This question is addressed next.

Comparing the Overlap Between the Two Lists

The data resulting from comparing the two sets of lists and the overlap between the OBW and CER lists as a whole is shown in Table 2. The data can be divided into three categories: (a) overlap at the same level, (b) overlap across the levels (with the preceding and succeeding levels), and (c) families that do not overlap, that is, those that occur in only one series. For example, the OBW Level 1 column shows that the 496 OBW word families at Level 1 occur at various levels of the CER lists. Sixteen OBW families at Level 1 do not overlap with the CER words at any level. The rows show the same kinds of data from a CER perspective. To provide a clearer picture of each kind of overlap, the data shown in Table 2 will be broken down into separate tables in the following sections. However, the reader will find it useful to keep referring back to Table 2 to see where the figures in the following tables came from.

The data in Table 2 was used to calculate the overlap of the two lists as a whole. The two series share 2,122 word families. All except 135 of the 2,257 families in the OBW list are in the CER list. From the OBW perspective, this is a 94.01% overlap, which is very large (see Figure 2).

Table 2. *Overlap between the OBW and CER wordlists for the new word families at each level*

	OBW level						Not in any of the OBW levels	Total
	1	2	3	4	5	6		
CER level								
1	377	67	13	5	1	1	13	477
2	80	144	44	18	14	7	13	320
3	17	77	110	41	32	5	57	339
4	4	19	83	109	92	56	139	502
5	1	2	36	73	171	124	263	670
6	1	1	7	19	82	189	448	747
Not in any of the CER levels	16	18	13	8	31	49		
Total	496	328	306	273	423	431		

From the CER perspective, the overlap is 69.46%. This smaller overlap results from the differing sizes of the two lists.

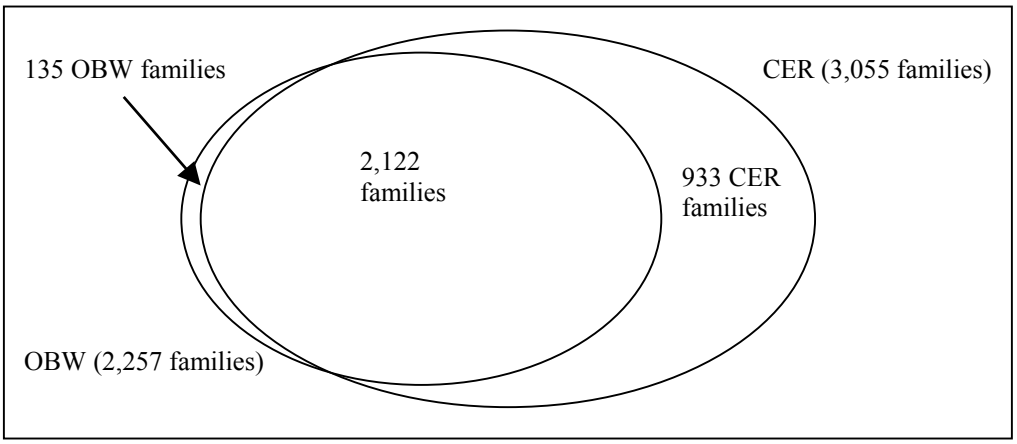


Figure 2. Overlap of the total new word families between the two series from the OBW perspective.

Comparing the Overlap of Total Word Families at Each Level

Overlap of word families at preceding levels plus current level families. The following analyses deal with the overlap of the actual word families occurring at each level of the OBW and CER schemes. First, the overlap of the families at each level is examined, for example, Level 1 of OBW with Level 1 of CER. This is one of the toughest tests of overlap: Level 2 includes the families at Level 1 plus those introduced at Level 2, Level 3 includes the families introduced at Level 3 plus all those of Levels 1 and 2, and so on. Next, the overlap between the families at each level of a series is looked at with the addition of the subsequent level. This is done because even if the overlap of families is not perfect at each level, the overlap may still be good because some of the overlapping families are at the next level of the series. Finally, the overlap of each level is compared with the current level and two subsequent levels.

The data in Table 3 is based on the cumulative overlap at and across the levels of the data in

Table 2. The figures were calculated in terms of total word families at the level. For example, the 668 families at Level 2 in Table 3 are the addition of the overlap at the preceding levels and overlap of Level 2 families. That is, 524 families (377 + 67 + 80) from the preceding levels (see Table 2) are added to 144 families as Level 2 overlap. This makes 668 families, which is the total overlap of OBW and CER at Level 2. The 929 families at Level 3 result from an addition of the overlap at the preceding levels (668 + 17 + 77 + 44 + 13 families) and the overlap at Level 3 (110 families), making a total of 929 families. The overlap of succeeding levels was calculated in the same way.

Table 3. *Overlap of families at preceding levels plus current level families from the OBW perspective*

	OBW level					
	1	2	3	4	5	6
	CER level					
	1	2	3	4	5	6
Number and percentage	377	668	929	1,208	1,630	2,122
of families overlapping	76.00%	81.06%	82.21%	86.10%	89.27%	94.01%

The percentage of the total number of word families shared by the two lists at each level from the OBW perspective is also shown in Table 3. To calculate the proportion, the total number of overlapping families at the level is divided by the total number of OBW families at that level. For example, 377 families are shared by OBW and CER at Level 1. From the OBW perspective, the 377 families are 76% of the 496 Level 1 OBW families (see Table 2) overlapping with Level 1 CER families. In a similar manner, from the OBW perspective, the 668 families at Level 2 are 81.06% of the Level 2 OBW families overlapping with Level 2 CER families. The 929 Level 3 OBW families are 82.21%, and so on.

From the OBW perspective, the figures indicating overlap at each level consistently increase, from 76% at Level 1 to 94.01% at Level 6. These figures show a sizable, but by no means perfect, cumulative overlap at each level.

Overlap of families at preceding levels plus current level families and families at the next level. The overlaps at the current level plus the next level are shown in Table 4, while the overlaps at the current level combined with the next two levels are shown in Table 5. To calculate these overlaps, the same steps of adding the overlapping families were taken as used for Table 3. Based on the data in Table 2, for example, the 457 families at Level 1 of Table 4 are the sum of 377 families (the previous overlap) and 80 families from the next level of CER. Then, 457 is divided by 496, which makes 92.14% at Level 1. The 1,320 families for Level 4 of Table 4 are the sum of 1,208 (the previous overlap) and 112 (1 + 2 + 36 + 73 as overlap at the next level of CER), which is then divided by 1,403, making 94.08% for Level 4. The proportions of the total overlap at the other levels were calculated in the same way. For Table 5, the overlap of the next two levels of CER was added to the total overlap at the level when proportions were calculated.

In Tables 4 and 5, we can see that the proportions of total overlap at every level are very high—well over 90% and close to 95%. A comparison of Tables 3 and 4 shows that most overlapping families are at the same level or the one following.

Table 4. *Overlap of word families at the preceding levels plus those at the current level and the next level from the OBW perspective*

	OBW level					
	1	2	3	4	5	6
	CER level					
	2	3	4	5	6	6
Number and percentage of families overlapping	457 92.14%	762 92.48%	1,035 91.59%	1,320 94.08%	1,740 95.29%	2,122 94.01%

Table 5. *Overlap of families at the preceding levels plus those at the current level and the next two levels from the OBW perspective*

	OBW level					
	1	2	3	4	5	6
	CER level					
	3	4	5	6	6	6
Number and percentage of families overlapping	474 95.56%	785 95.27%	1,074 95.04%	1,348 96.08%	1,740 95.29%	2,122 94.01%

The following tables contain the figures calculated from data in Table 2, but from the CER perspective. The same steps for calculating the number of families were applied to the CER lists. The results in terms of the overlap at the preceding levels and current level families are shown in Table 6, while more detail about the overlap at each level plus families at the next levels is given in Tables 7 and 8.

Table 6. *Overlap of word families at preceding levels plus current level families from the CER perspective*

	CER level					
	1	2	3	4	5	6
	OBW level					
	1	2	3	4	5	6
Number and percentage of families overlapping	377 79.03%	668 83.81%	929 81.78%	1,208 73.74%	1,630 70.62%	2,122 69.46%

The overlap of families at the preceding levels plus the current level families from the CER perspective is considerable at the three lower levels of the CER series and is less at the three higher levels, as shown in Table 6. This is because of the greater numbers of word families at these levels compared with OBW. The overlap ranges from 69.46 to 83.81%. When added to the overlap at the next level, as seen in Table 7, the overlap at most levels increases, particularly at the three lower levels, to around 90%. The same pattern is seen in Table 8, where the current level plus the next two levels are considered.

Particularly from an OBW perspective, but to a large degree also from the CER perspective, the two lists have a considerable degree of overlap. The differences are largely the results of differences in the sizes of the two lists, rather than in the actual families in the lists or the sequencing of these families into levels. This is reassuring for users of graded readers, indicating that the two series of readers have similarities in vocabulary grading.

Table 7. *Overlap of word families at the preceding levels plus those at the current level and at the next level from the CER perspective*

	CER level					
	1	2	3	4	5	6
	OBW level					
	2	3	4	5	6	6
Number and percentage of families overlapping	444 93.08%	725 90.97%	993 87.41%	1,347 82.23%	1,823 78.99%	2,122 69.46%

Table 8. *Overlap of word families at the preceding levels plus those at the current level and at the next two levels from the CER perspective*

	CER level					
	1	2	3	4	5	6
	OBW level					
	3	4	5	6	6	6
Number and percentage of families overlapping	457 95.80%	748 93.85%	1,040 91.55%	1,416 86.45%	1,823 78.99%	2,122 69.46%

Overlap Between the GSL *Words and the OBW and CER Lists*

The most well-known general-service-vocabulary list is the GSL, and it has been the basis for many series of graded readers. How similar are the lists used in graded readers and the GSL?

Table 9 shows that 360 word families of Level 1 in the OBW series are in the 1,000-word level of the GSL. Level 2 has 182 families in 1,000-word level. A total of 921 out of the 990 families in the 1,000-word level are in the OBW.

Table 9. *The 1,000- and 2,000-word levels of the GSL in the OBW lists*

	OBW level						GSL families in the OBW lists	GSL families not in the OBW lists	Total GSL families
	1	2	3	4	5	6			
GSL families									
1st 1,000	360	182	131	87	90	71	921 93.03%	69	990
2nd 1,000	77	105	110	121	164	164	741 76.07%	233	974
Total	437	287	241	208	254	235	1,662 84.62%	302	1,964

Sixty-nine word families are in the 1,000-word level of the GSL but not in the OBW lists. These include words like *arise*, *affair*, *base*, *entire*, and *favour.* The overlap with the second 1,000 of the GSL is not as good, and the total overlap of the GSL and OBW is 84.62%, with 302 families in the GSL but not in OBW. The overlap between the GSL and CER is higher than that between the GSL and OBW, but this is largely because the CER list contains over 1,000 more families than the GSL.

As shown in Tables 9 and 10, the proportions of the first 1,000 GSL families included in the two lists are high: 93.03% in the OBW list and 96.16% in the CER list. For the second 1,000 families, between the two series, the degree of overlap is lower: 741 families in the OBW list and 802 families in the CER list.

Table 10. *The 1,000- and 2,000-word levels of the GSL in the CER lists*

	CER level						GSL families in the CER lists	GSL families not in the CER lists	Total GSL families
	1	2	3	4	5	6			
GSL families									
1st 1,000	324	191	146	116	118	57	952 96.16%	38	990
2nd 1,000	94	84	89	149	185	201	802 82.34%	172	974
Total	418	275	235	265	303	258	1,754 89.31%	210	1,964

Tables 9 and 10 show that 69 of the first 1,000 GSL families are not in the OBW list and that 38 of the first 1,000 GSL families do not appear in the CER list. However, 20 of these families overlap, so 87 of the families from the 1,000-word level of the GSL are not in the two publishers' lists. These 20 families were *association*, *English*, *forth*, *form*, *hurrah*, *mass*, *mere*, *ounce*, *regard*, *scale*, *stock*, *base*, *difference*, *honour*, *native*, *production*, *poverty*, *standard*, *subject*, and *upon*. The same pattern occurred with the second 1,000 GSL families in the two lists. The total number of GSL families that are not in the OBW and the CER lists is shown in Figures 3 and 4.

We can see in Figures 3 and 4 that the number of GSL families not occurring in the two lists is not large. For the first 1,000 GSL words, the total number is 87, which is 8.79% of 990 families, while for the second 1,000 GSL words, the total is 280, or 28.74% of 974 families. In total, 367 of the 1,964 families (18.85%) in the GSL do not occur in two series of readers. From the GSL perspective, the difference is not large between the two lists and the GSL. This perspective is preferable because the GSL (1,964 families) is smaller than the OBW lists (2,257 families) and the CER lists (3,055 families).

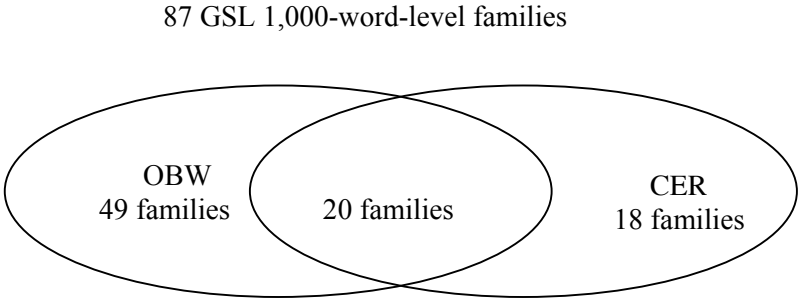


Figure 3. The families of the GSL 1,000-word level that do not occur in the OBW and the CER lists.

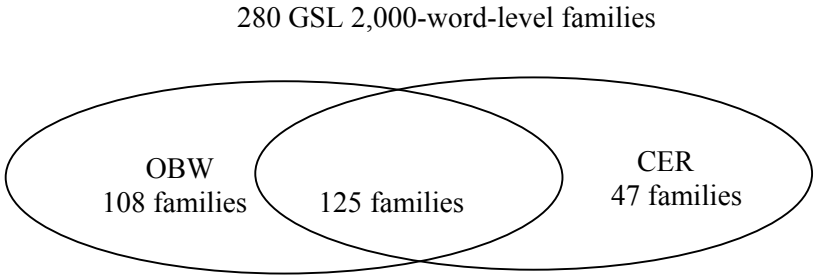


Figure 4. The families of the GSL 2,000-word level that do not occur in the OBW and the CER lists.

The numbers of OBW and CER families not in the GSL are shown in Tables 11 and 12. This is a less preferable perspective to see the overlap of the OBW and CER lists. Because both lists are larger than the GSL, the number of families not occurring in the GSL is large. The number of families not in the GSL increases in both series from the lower to upper levels. The number of OBW and CER families not in the GSL is small at the three lower levels. That is, from 41 to 65 families not in the GSL are found in the four lower levels of the OBW, and the three lower levels of the CER range from 45 to 104 families not in the GSL. At the two upper levels of the OBW and the three upper levels of CER, the numbers of families in the two lists increase. Because the CER introduces more families in the lists (3,055 families), this results in a large number of families not in the GSL. In total, the CER includes 1,301 families not in the GSL, while the OBW (2,257 families) includes 595 families not in the GSL.

Table 11. *Number of OBW families not in the* GSL

OBW level	GSL			OBW subtotal	Families not in the GSL
	1st 1,000	2nd 1,000	Total		
1	360	77	437	496	59
2	182	105	287	328	41
3	131	110	241	306	65
4	87	121	208	273	65
5	90	164	254	423	169
6	71	164	235	431	196
Total	921	741	1,662	2,257	595

Table 12. *Number of CER families not in the* GSL

CER level	GSL			CER subtotal	Families not in the GSL
	1st 1,000	2nd 1,000	Total		
1	324	94	418	477	59
2	191	84	275	320	45
3	146	89	235	339	104
4	116	149	265	502	237
5	118	185	303	670	367
6	57	201	258	747	489
Total	952	802	1,754	3,055	1,301

A comparison of the families not in the GSL in the two lists showed that 399 families occurred in

both series. The overlap between the two series in terms of the total number of the families not in the GSL is illustrated in Figure 5.

Overall, the data suggests that because of the size differences between the GSL and the two graded-reader series, particularly the CER list, none of them could representatively cover the families in the two graded-reader series, particularly at the higher levels of the series.

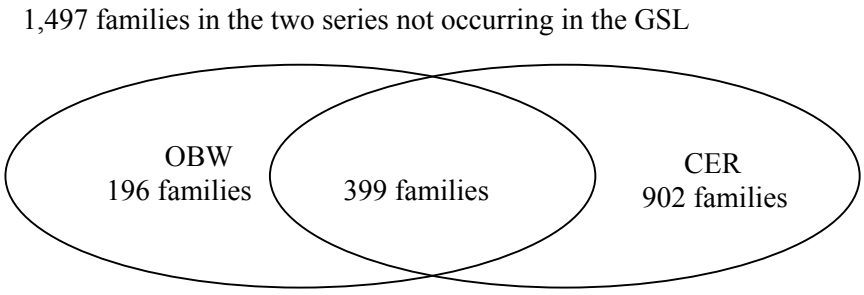


Figure 5. Number of the OBW and CER families not occurring in the GSL.

Do the Books Follow the Publishers' Lists?

The available lists have been compared to see how similar they are. If the actual books do not follow these lists, then the comparison is meaningless. The next step in the study, therefore, was to see how closely the vocabulary in the books resembled that in the lists. Two books were chosen to represent the books at each of three levels: Levels 1, 3, and 5. The books were scanned and used as input texts to be analyzed by the Range program using the OBW and CER baseword lists. The overlap at each level is the number of families that occur in both the books and the publishers' lists.

Then, from the OBW perspective, books at Levels 1, 3, and 5 from three series, CER, Penguin Readers (PR), and Macmillan Guided Readers (MGR), were analyzed to see what coverage the OBW list gave of these books.

Families actually used in the books from the OBW perspective. The books from the three levels of OBW are *White Death* (Book 1of Level 1), *The Lottery Winner* (Book 2 of Level 1), *The Picture of Dorian Gray* (Book 1 of Level 3), *Ethan Frome* (Book 2 of Level 3), *The Dead of Jericho* (Book 1 of Level 5), and *The Garden Party* (Book 2 of Level 5). They were all examined in the same way. Data detailing the word families actually used at each level are shown in Table 13. This first section looks at the actual word families used. The next section looks at the coverage.

In the first book of OBW Level 1 (i.e., *White Death*), 303 Level 1 word families are used, and the book contains a total of 361 word families, as shown in Table 13. Therefore, the proportion of Level 1 word families actually used in *White Death* is 83.93% (303 divided by 361). When the number of the proper nouns is added to this proportion (83.93% + 4.70%), it equals 88.63% (see the "Families and proper nouns" column). The book also uses 22 word families from Level 2 of the OBW scheme, 6 from Level 3, and so on. The book uses 6 families that are not on the OBW list. Another Level 1 OBW book, *The Lottery Winner*, uses 84.63% Level 1 families and proper

nouns. The overlap between the list and the vocabulary in the text is not nearly as high as it should be in both books.

A larger amount of overlap is found in the two upper levels. At Level 3, the two OBW books, *The Picture of Dorian Gray* and *Ethan Frome*, have overlaps of 94.54 and 90.46% when proper nouns are included. The two Level 5 OBW books also provide a rather high overlap of the families at Level 5, that is, 92.26 and 92.88%. The details are given in Table 13. In terms of actual families used in a book, the bigger the overlap, the better the book.

Table 13. *Overlap of the word families actually used in the OBW books at each level from the OBW perspective*

	OBW level						Proper	Families and	Total word	Words not in
OBW book	1	2	3	4	5	6	nouns	proper nouns	families	any levels
B1-1	303 (83.93)	22	6	4	2	1	17 (4.70)	320 (88.63)	361	6
B1-2	334 (76.61)	36	6	7	1	1	35 (8.02)	369 (84.63)	436	16
B3-1	342	178	118	9	4	4	38	676	715	22
	638 (89.23)						(5.31)	(94.54)		
B3-2	343	210	136	13	11	4	70	759	839	52
	689 (82.12)						(8.34)	(90.46)		
B5-1	409	261	199	129	128	13	140	1,266	1,372	93
	1,126 (82.06)						(10.20)	(92.26)		
B5-2	390	241	175	134	139	23	69	1,148	1,236	69
	1,079 (87.30)						(5.58)	(92.88)		

Note. The values in parentheses are percentages. B1-1 = Book 1 of Level 1; B1-2 = Book 2 of Level 1; B3-1 = Book 1 of Level 3; B3-2 = Book 2 of Level 3; B5-1 = Book 1 of Level 5; B5-2 = Book 2 of Level 5.

In *White Death*, 303 Level 1 families were used from the available 400 headwords in the Level 1 OBW list. This is 75.75%. *The Lottery Winner* used 83.5% of the 400 Level 1 families in the list. Because 95 or 98% coverage is needed, more Level 1 words should be used in the two Level 1 books to reduce the heavy vocabulary load of unknown words.

White Death uses 6 words not in any OBW levels: *bedroom*, *courtroom*, *jury*, *toothpaste*, *tube*, and *tubes*, most of which were related to the story and made the story real to its audience when the storyline was presented. The word *courtroom* occurred 13 times, and the word *jury* appeared 19 times in the book. Such occurrences may help the readers increase their knowledge of the two words and decrease the burden of unknown words when the readers work through the book. In a similar manner, readers would encounter the word *tubes* 22 times in the book, and may thus be able to get the meaning during reading. As for *bedroom* (1 occurrence) and *toothpaste* (34 occurrences), single stems like *bed*, *room*, and *tooth* can help the reader guess their meanings more easily.

The second major group of words used in *White Death* is 22 Level 2 words. Among those words are *police* (30 occurrences), *inspector* (24 occurrences), and *prison* (12 occurrences). Each word is essential for the story. Although these words are beyond the current level, they are repeated

often and do not continue to be burdens.

The Lottery Winner contains 16 words not in any levels. They are *cell* (3 occurrences), *champagne* (5 occurrences), *charity* (4 occurrences), *footballer* (1 occurrence), *lotteries* (10 occurrences), *lottery* (44 occurrences), *interestingly* (1 occurrence), *snatched* (6 occurrences), *snatcher* (3 occurrences), *snatching* (1 occurrence), *stage* (1 occurrence), *sunshine* (10 occurrences), *huh* (1 occurrence), *eh* (1 occurrence), *mmm* (1 occurrence), and *ah* (1 occurrence). The word *lottery*, the topic word, clearly occurs the most throughout the book, and the word *snatched* is also clearly important in the story. The 36 words of Level 2, including *policeman* (24 occurrences), *winning* (17 occurrences), and *stole* (11 occurrences), relate to the topic presented in the story. The words that occur only once are the ones that indicate poor vocabulary control, and the words occurring more than once or twice are unlikely to detract from the accessibility of the text. In addition, a high number of repetitions of unknown words can help the learner guess their meanings in the story.

Words actually used in the books from the CER perspective. The same method as used with the OBW books was applied to six CER books: *Inspector Logan* (Book 1of Level 1), *Parallel* (Book 2 of Level 1), *The House by the Sea* (Book 1 of Level 3), *The Ironing Man* (Book 2 of Level 3), *All I Want* (Book 1 of Level 5), and *The Emergency Murder* (Book 2 of Level 5).

The overlap of the CER lists and the six CER books from the CER perspective is shown in Table 14. At Level 1, in *Inspector Logan*, 280 of the 372 families are Level 1 CER families. When added to the proper nouns (20), they provide 80.65% coverage (see the "Families and proper nouns" column), while the other Level 1 CER book gives 77.74% coverage. The two Level 3 CER books overlap with the lists 88.39 and 86.91%, and the two Level 5 CER books overlap 88.76 and 91.73%.

Table 14. *Overlap of word families actually used in the CER books at each level from the CER perspective*

	CER level						Proper	Families and	Total word	Words not in
CER book	1	2	3	4	5	6	nouns	proper nouns	families	any levels
B1-1	280 (75.27)	39	10	9	2	2	20 (5.38)	300 (80.65)	372	10
B1-2	249 (71.96)	58	6	5	4	1	20 (5.78)	267 (77.74)	346	3
B3-1	342	197	109	43	15	4	30	678	767	27
	648 (84.48)						(3.91)	(88.39)		
B3-2	365	218	144	48	15	8	50	777	894	46
	727 (81.32)						(5.59)	(86.91)		
B5-1	361	221	159	142	107	31	92	1,082	1,219	106
	990 (81.21)						(7.55)	(88.76)		
B5-2	399	254	187	208	162	13	90	1,310	1,428	115
	1,210 (85.43)						(6.30)	(91.73)		

Note. The values in parentheses are percentages. B1-1 = Book 1 of Level 1; B1-2 = Book 2 of Level 1; B3-1 = Book 1 of Level 3; B3-2 = Book 2 of Level 3; B5-1 = Book 1 of Level 5; B5-2 = Book 2 of Level 5.

A comparison of the words actually used in the two Level 1 CER books with the families in the Level 1 CER list showed that the first book uses 70% of the 400 headwords of the Level 1 CER list and that the other Level 1 book uses 64% of the Level 1 words. *Inspector Logan* and *Parallel* do not provide learners with a large proportion of the current level words. Many of the words used in the two Level 1 CER books are Level 2 words, that is, 39 and 58 words.

In *Inspector Logan*, the 10 words not in any levels include *lunchtime* (1 occurrence), *questioningly* (1 occurrence), *rental* (2 occurrences), *teabags* (1 occurrence), *tomorrows* (1 occurrence), *scientists* (2 occurrences), *sergeants* (9 occurrences), *sergeant* (1 occurrence), *somethings* (1 occurrence), and *terrace* (7 occurrences). Most of the off-list words occur once except for the words *sergeants* and *terrace*, which are important in the story. Repetitions of the two words may help learners recall their meanings when working through the story. In the other Level 1 book, *Parallel*, the 3 words not in any levels are *chin* (1 occurrence), *knees* (2 occurrences), and *prologue* (1 occurrence). The word *prologue* is used to introduce the book before the story was presented. With very few occurrences, these words are unlikely to affect an overall understanding of the story.

Level 2 words are the second major group of words used in the two Level 1 CER books as in the two Level 1 OBW books. Among the 39 Level 2 words in *Inspector Logan*, the word *killed*, which is related to the topic, occurs 9 times throughout the book. Several of the Level 2 CER words in *Parallel*, such as *different* (25 occurrences) and *around* (10 occurrences), are not obviously topic related.

The data indicates the overlap in terms of word families. A look at the number of tokens in a whole book will reveal a clearer picture of the coverage of texts, which is expected to assist learners in coping with reading. Here, *token* refers to each occurrence of a word that is counted each time it occurs in the text, and *coverage* refers to the percentage of the tokens in a text or corpus covered by a particular word list. A high coverage indicates that vocabulary may not be a problem in reading the text because the unknown words are few within a largely known context. A 95% coverage is a good start (Laufer, 1989; Liu and Nation, 1985), but a 98–100% coverage is preferable for graded readers (Nation, 2001).

Text coverage of the OBW books at each level from the OBW perspective. The six OBW books provide a reasonable coverage from the OBW perspective, as shown in Table 15. For example, in *White Death*, 303 OBW Level 1 word families occurred in the book as 6,035 tokens. Then, 6,035 divided by 6,869 running words in the book makes 87.86% coverage of the text. When the occurrences of proper nouns (6.39%) in the book are added, the coverage of the text at Level 1 is 94.25% (see the "Tokens and proper nouns" column). This proportion is close to 95% coverage, which is expected to enhance guessing from context when some difficult words are introduced. As we can see in Table 15, the four books at Levels 3 and 5 give good coverage at their levels. They cover 97.07 and 97.34% at Level 3 and 98.60 and 98.59% at Level 5. These figures indicate that the vocabulary is not controlled as well at Level 1 in the series as it is in the later levels. That is, Level 1 books have more words outside the level than books at Levels 3 and 5.

Text coverage of the CER books at each level from the CER perspective. A similar pattern for the six CER books is shown in Table 16. In *Inspector Logan*, for example, the coverage of the Level

1 text is 92.05% (see the "Tokens and proper nouns" column). This is the result of 3,480 occurrences of Level 1 CER words (280 word families) plus 345 occurrences of proper nouns (20 words) throughout the book, which is divided by the total number of running words in the book (4,155 tokens). The other Level 1 book has 91.55% coverage. These are well below the coverage needed for unassisted reading. The percentage coverage is higher at the upper levels, as found in Levels 3 and 5. It is 98.39 and 97.63% at Level 3 and 96.58 and 98.59% at Level 5.

Table 15. *Token coverage of the OBW books at each level from the OBW perspective*

OBW book	OBW level						Proper nouns	Tokens and proper nouns	Total tokens	Tokens not in any levels
	1	2	3	4	5	6				
B1-1	6,035 (87.86)	167	29	57	23	17	439 (6.39)	6,474 (94.25)	6,869	102 (1.48)
B1-2	4,994 (86.72)	210	20	74	3	23	337 (5.85)	5,331 (92.57)	5,759	98 (1.70)
B3-1	8,319 (77.33)	993 (9.23)	477 (4.43)	54	31	36	654 (6.08)	10,443 (97.07)	10,758	194 (1.80)
B3-2	8,297 (75.51)	1,254 (11.41)	543 (4.94)	57	57	17	602 (5.48)	10,696 (97.34)	10,988	161 (1.46)
B5-1	18,037 (78.59)	2,054 (8.95)	969 (4.22)	485 (2.11)	317 (1.38)	52	768 (3.35)	22,630 (98.60)	22,952	270 (1.17)
B5-2	18,373 (74.98)	2,178 (8.89)	1,087 (4.44)	537 (2.19)	483 (1.97)	98	1,500 (6.12)	24,158 (98.59)	24,503	247 (1.01)

Note. The values in parentheses are percentages. B1-1 = Book 1 of Level 1; B1-2 = Book 2 of Level 1; B3-1 = Book 1 of Level 3; B3-2 = Book 2 of Level 3; B5-1 = Book 1 of Level 5; B5-2 = Book 2 of Level 5.

Table 16. *Token coverage of the CER books at each level from the CER perspective*

CER book	CER level						Proper nouns	Tokens and proper nouns	Total tokens	Tokens not in any levels
	1	2	3	4	5	6				
B1-1	3,480 (83.75)	133	28	38	11	87	345 (8.30)	3,825 (92.05)	4,155	33 (0.79)
B1-2	3,645 (85.66)	236	34	36	32	17	251 (5.89)	3,896 (91.55)	4,255	4 (0.09)
B3-1	13,403 (82.03)	1,619 (9.94)	516 (3.17)	72	26	26	530 (3.25)	16,068 (98.39)	16,286	94 (0.57)
B3-2	11,633 (78.39)	1,721 (11.54)	663 (4.47)	145	26	14	480 (3.23)	14,497 (97.63)	14,840	158 (1.06)
B5-1	14,707 (70.48)	1,845 (8.84)	773 (3.70)	548 (2.63)	329 (1.58)	263	1,952 (9.35)	20,154 (96.58)	20,866	449 (2.15)
B5-2	19,048 (75.22)	2,859 (11.29)	1,062 (4.19)	715 (2.82)	369 (1.46)	78	915 (3.61)	24,968 (98.59)	25,322	276 (1.08)

Note. The values in parentheses are percentages. B1-1 = Book 1 of Level 1; B1-2 = Book 2 of Level 1; B3-1 = Book 1 of Level 3; B3-2 = Book 2 of Level 3; B5-1 = Book 1 of Level 5; B5-2 = Book 2 of Level 5.

The books of the two series are likely to give reasonable coverage at the higher levels, as shown in Tables 15 and 16. However, at the lower levels, the Level 1 words and proper nouns cover

only 92.60% on average, which is not close enough to 95 or 98% coverage.

To further examine the coverage in various graded-reader series, books of three series, CER, PR, and MGR, were chosen. They were analyzed from the OBW perspective (using the OBW baseword lists). The OBW lists were chosen as a basis for investigating the books for several reasons. First, evidence from a descriptive analysis of the wordlists of OBW and CER shows that 94.01% of the families in the OBW list are in the CER list. This ensures that the OBW itself can represent a large number of families shared by the CER, while the CER cannot, because of its much larger size. When the size of the OBW list is compared with those of the PR and the MGR (using the vocabulary size in the catalogues because the lists are not given to the public), the differences are not large: 2,200 (MGR), 2,500 (OBW), and 3,000 (PR). Second, if the number of levels is an issue, the OBW has six levels, as do the CER and the PR, although the MGR has only five levels. The last reason is that the OBW graded-reading scheme provides good coverage at higher levels for its books.

Initially, the OBW series was expected to be a good basis for investigating token coverage of other series. The question of whether this is true will be addressed next.

To look at the effect of using the OBW lists as a standard, the words actually used in the CER books were reexamined from the OBW perspective. The same method was applied to six books of the PR series: *The Missing Coins* (Book 1of Level 1), *The House of the Seven Gables* (Book 2 of Level 1), *The Yearling* (Book 1 of Level 3), *The Hunchback of Notre-Dame* (Book 2 of Level 3), *Prime Suspect* (Book 1 of Level 5), and *The Warden* (Book 2 of Level 5).

Similarly, six books of the MGR series were chosen to represent the three levels. They are *Alissa* (Book 1of Level 1), *Paradise Island* (Book 2 of Level 1), *The Runaways* (Book 1 of Level 3), *The Black Cat* (Book 2 of Level 3), *Great Expectations* (Book 1 of Level 5), and *The Man of Property* (Book 2 of Level 5).

To give a clearer picture of the sample books of the four series, Table 17 summarizes the text coverage of the books at the three levels studied.

Table 17. *Percentage text coverage of the books of four graded-reading schemes from the OBW perspective*

Series	Level 1		Level 3		Level 5	
	Book 1	Book 2	Book 1	Book 2	Book 1	Book 2
OBW	94.25	92.57	97.07	97.34	98.60	98.59
CER	92.15	90.45	97.90	96.05	95.52	97.34
PR	89.21	93.82	96.39	94.78	97.45	94.76
MGR	87.16	78.24	92.76	96.04	95.87	97.66

Generally speaking, the OBW books come out best when analyzed using the OBW list. The two Level 5 OBW texts are the only ones to reach the desired 98% coverage. In many cases, however, the differences in coverage between the books from the other series and the OBW books are not great.

The coverage of the Level 1 books in all the series including OBW is not satisfactory. The coverage figures are all below 95%, for several possible reasons. First, writing books at this level may not be possible using such a limited vocabulary. Second, the lists are not well made and thus cannot do the job that they are supposed to do. Finally, writers and editors may not be applying the lists very stringently, letting the story determine the words.

Discussion

An analysis and comparison of the lists of the word families in the two major series, OBW and CER, was carried out using the Range program. The results gave information about the wordlists of the graded readers and furthered understanding of the relationship between them and the GSL words.

The results of the study indicate that the wordlists of the graded readers exploit high-frequency words to provide readable texts suitable for establishing known vocabulary and learning unknown vocabulary. In the three lower levels of the OBW and CER series, the number of word families in the two lists does not differ much, and the series overlap considerably.

At the higher levels of the series (Levels 4 to 6) are considerable size differences, and as a result, only a small amount of overlap between the series for the new word families introduced at those levels. Although the 2,000 GSL families are crucial for learners of English, the OBW and the CER contain more families than does the GSL. The OBW has 595 more, and the CER has 1,301 more. In addition, the words actually used in the books do not stick closely enough to the families on the lists on which they are based, especially at Level 1.

The study may provide useful information for teachers using graded readers in extensive reading programs. The findings highlight that the number of headwords cannot be used as a good criterion for comparing the series. For categorizing books in extensive reading, the findings based on the size differences between the lists of the two series suggest that when the books of various series need to be shelved or categorized in a reading room, they should be classified separately in their own series. They cannot be sensibly compared simply on the basis of similar numbers of headwords or levels. In addition, in practice, when using the graded-reading schemes of various series, the differences in vocabulary sizes, divisions of levels, and actual words used make it impossible to take one scheme as a good representative of the others and in this way develop a categorization that will fit all schemes. This suggests that teachers need to be very flexible when setting up standard sets of levels for a graded-reading library incorporating books from various series.

In general, the books do not stick closely to the words in the publishers' lists, particularly at Level 1. One cause of this is that words that are important for a story are brought in, and some of these words are repeated throughout the text. This is acceptable, and these words are likely to be learned without becoming burdens for the readers. When words outside the books and the lists are brought in and used only one or two times, this is evidence of careless simplification. However, coverage from the OBW perspective suggests that the graded-reading schemes of each series have attempted to provide suitable conditions for unassisted reading, particularly at higher levels. This suggests that reading across the series is possible if learners have gained enough

knowledge of the high-frequency words needed to read at that level. From a practical perspective, teachers may have difficulty knowing whether their learners know sufficient high-frequency words for each level. Teachers should be aware of the different words used in the books of different series when assigning learners to read any series of graded readers. Without careful attention to the different numbers of headwords in the books of different series, reading across the series might affect learners' reading abilities and lead to unfavorable attitudes towards the habit of reading. In sum, to promote the learning of words in graded readers, learners should be assigned to read and work their way through the levels within one graded-reading scheme at a time.

The difference in the sizes of the lists and actual words used in the books reminds teachers that learners will meet quite a high proportion of different unknown words when they move to a new level in their graded reading. If the idea of guessing from context is applied, because a considerable proportion of the GSL words are included in the graded-reader lists, supplementing the learning through reading with direct study of the GSL words would be wise. This helps increase the learners' knowledge of the essential high-frequency vocabulary required for the learning of the unknown words in context when learners are engaged in reading various levels of graded readers. At this point, word cards would be best used individually (Nation, 2001), with learners making their own cards and choosing the words from the GSL lists. By doing so, the density of unknown words will become light enough to allow more fluent reading. This should enhance reading across the various series.

The difference in vocabulary sizes and divisions of levels found in the study also suggests that extensive reading, particularly that done with graded readers, should be assigned with great care. This is to control ability levels. If too many words are unknown and learners lack motivation, they will not make many gains. Despite graded readers being used as simplified texts to increase both the learning of words and fluency in reading, learners are probably unable to take advantage of being exposed to more unknown language below a certain vocabulary threshold. If learners get too much new input and it is not comprehensible, their gains are likely to be few. Conversely, without new input, their chances to learn and demonstrate learning will be few. In other words, learners who are given materials that are too easy are not challenged, and their growth can be hampered (Chall & Conard, 1991). The findings clearly address a vocabulary size issue and suggest that teachers should place learners in appropriate levels of reading to reach the ultimate goal of extensive reading with graded readers.

In terms of placing learners at appropriate levels of graded reading, if a placement test is needed, none of the three lists can act as a good source of words for testing for all series. Because of the size differences between the GSL and the two graded-reader series, particularly the CER lists, sampling words from the GSL for a placement test for extensive reading would not give a representative coverage of words in the graded-reader series, particularly at the higher levels of the series. If a test based on the GSL was used for the first four levels of the two series, it would be reasonably representative, but a substantial number of words would still be in the GSL and not in Levels 1 to 4, and a substantial number of words would still be in the two series and not in the GSL. The size differences are large between the two wordlists and the GSL. That is, the two series contain more words than are in the GSL. A total of 1,497 words occur in the two series but not in the GSL. This evidence does not support the idea of making a test from the GSL words to

represent the words in the graded readers. The GSL is not a feasible source of words for a placement test for extensive reading. The findings of this study answer the question of whether words from the wordlists of graded readers and the GSL could be a good source of words for a vocabulary size test for graded-reading schemes. They suggest that using the GSL as a source of words for a placement test is problematic. Moreover, as the two series differ from each other considerably in size, beyond the first three levels of the series, expecting a vocabulary measure to properly represent the words across the two series is clearly not feasible. Teachers or researchers who are looking for a word source for making such tests should look elsewhere.

This study has worked within a narrow focus. It has looked only at the wordlists of graded readers and the GSL in terms of a word source for the learning of words in simplified texts. This study has attempted to examine similarities and differences in the sizes of lists and actual words used in the books, which provides an understanding of the wordlists in detail. Most graded readers are designed according to their own wordlists and deliberately not set up as a way of presenting new vocabulary; rather, they are seen as being supplementary readers that help establish vocabulary already met in language courses. This is in line with the results of the Nation and Wang study (1999).

Despite the limitations in the size of the study, the findings add to our understanding of the wordlists of graded readers and their relationship with the GSL and actual words used in the books. An interesting issue for further study would be to develop the teaching and learning of vocabulary through extensive reading with graded readers when control of ability is needed. For example, how can we measure a vocabulary size that may enhance reading across series? What measures can be assigned to test the learning of words from graded readers? Is a placement test for graded-reading schemes feasible?

Acknowledgments

I am grateful to Professor Paul Nation, my PhD advisor, at Victoria University of Wellington, for his generous help with access to the OBW and CER wordlists and for his comments on an earlier draft of this article. I also thank the two publishers and the anonymous reviewers who provided valuable comments to clarify this article. This research was supported by a grant from the Faculty of Humanities and Social Sciences, Mahasarakham University, Thailand. Without all mentioned, this study would never have been completed.

Notes

1. The GSL, developed in the 1940s, contains 2,000 headwords. The frequency figures for most items are based on a 5 million-word written corpus. Percentage figures are given for different meanings and parts of speech of the headwords. In spite of its age, occasional errors, and solely written base, it remains the best of the available lists not only because of its information about the frequencies of meanings but also because of West's careful application of criteria other than frequency and range. The 2,000 GSL words are of practical use to teachers and curriculum planners because they are contained within word families, each with its own frequency.

2. Coverage refers to the percentage of the tokens in a text or corpus contained in a particular word list. Text coverage helps readers guess from context and build fluency in reading by providing good proportions of known words. See more details in Schmitt and McCarthy (1997, pp. 6–19).

3. The AWL contains 570 word families and does not include words that are in the most frequent 2,000 word families of English. For information on the development and evaluation of the AWL, see Coxhead (2000).

References

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*, 253–279.

Chall, J. S., & Conard, S. S. (1991). *Should textbooks challenge students? The case for easier or harder books*. New York: Teacher College Press.

Chung, T. M., & Nation, I. S. P. (2004). Identifying technical vocabulary. *System*, *32*, 251–263.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

Hill, D. (1997). Survey review: Graded readers. *English Language Teaching Journal*, *51*(1), 57–81.

Hill, D. (2001). Graded readers. *English Language Teaching Journal*, *55*, 300–324.

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*, 689–696.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430.

Hwang, K., & Nation, I. S. P. (1989). Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. *Reading in a Foreign Language*, *6*, 323–335.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordmann (Eds.), *From humans thinking to thinking machines* (pp. 316–323). Clevedon: Multilingual Matters.

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.

Liu, N., & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, *16*, 33–42.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–14). Amsterdam: John Benjamins.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82.

Nation, I. S. P., & Wang, M. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, *12*, 355–380.

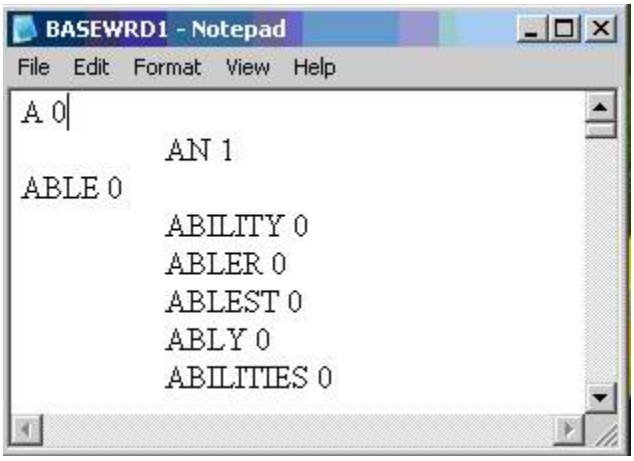
Nation, I. S. P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [Computer software]. Retrieved from http://www.vuw.ac.nz/lals/staff/paul-

nation/nation.aspx
Schmitt, N., & McCarthy, M. (Eds.). (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge, UK: Cambridge University Press.
Sutarsyah, C., Nation, I. S. P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal*, *25*, 34–50.
West, M. (1953). *A general service list of English words*. London: Longman.

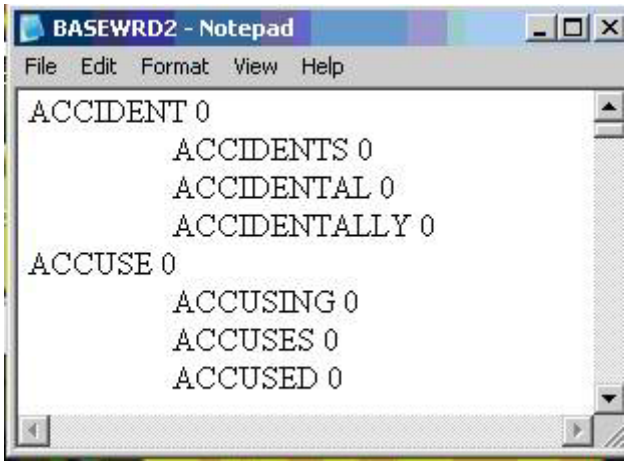
Appendix A

Three Baseword Lists

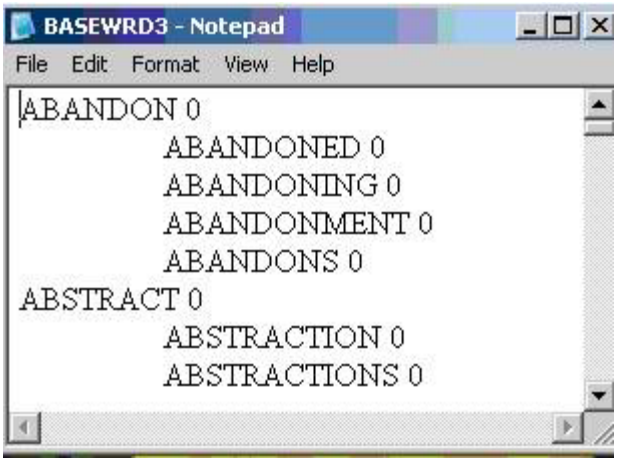
Some examples of the word families with their family members in the baseword lists for the Range program are shown below. The program compares the word forms in texts with three baseword lists built into the program consisting of the first 1,000 and the second 1,000 families of the GSL and the AWL.



The first baseword list consists of the first 1,000 word families.



The second baseword list consists of the second 1,000 word families

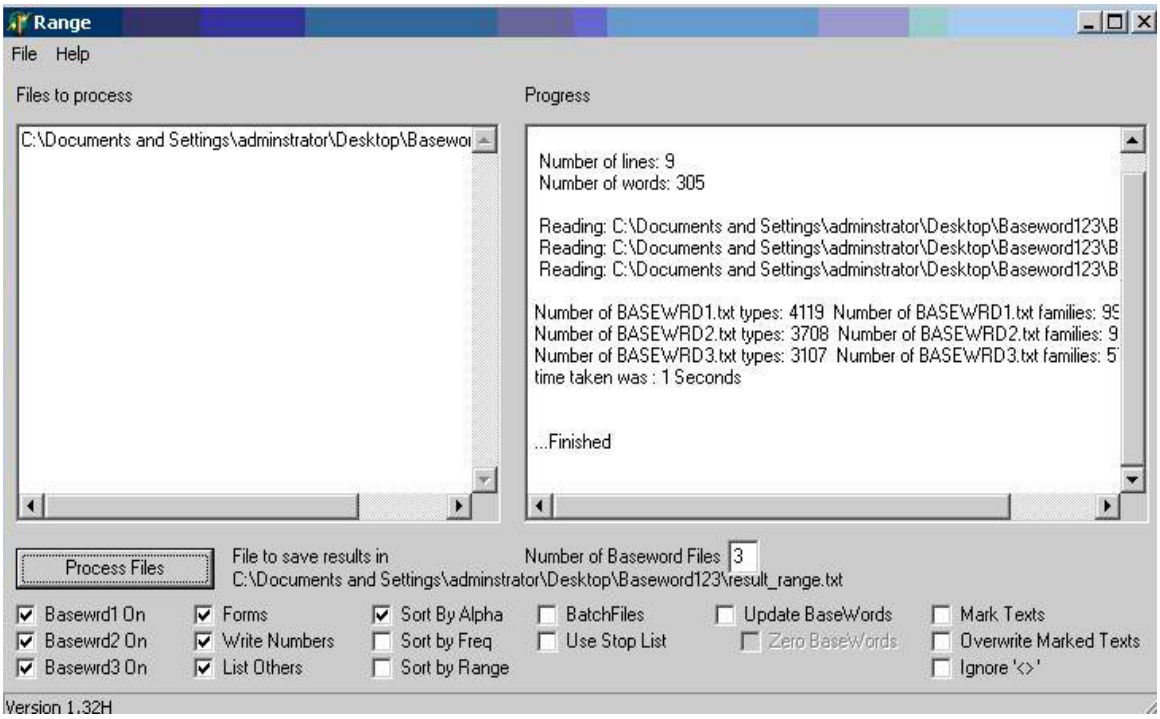


The third baseword list consists of academic word families

Appendix B

Samples of the Output of the Range Program

The following screenshot illustrates the output of the Range program using the three baseword lists and a short text taken from Level 1 of the OBW series.



The box on the right shows the three baseword lists and a processed file produced when a text (from Level 1 of the OBW) was input to the Range program.

The following output shows how many word families in the input text were found in each list. For example, 109 word families were in the first list, 21 in the second list, none in the AWL, and 6 words were outside the word lists. In addition, text coverage can be determined to see whether this text supports

reading ability at a certain level. This text had 84.59% coverage, which means that it is likely to contain many Level 2 words, unknown at Level 1, and this could affect reading ability at Level 1 if comprehension and fluency in reading is required.

> Processing file: C:\Documents and Settings\adminstrator\Desktop\Baseword123\OBW 1.txt
> Number of lines: 9
> Number of words: 305
>
> Reading: C:\Documents and Settings\adminstrator\Desktop\Baseword123\BASEWRD1.txt
> Reading: C:\Documents and Settings\adminstrator\Desktop\Baseword123\BASEWRD2.txt
> Reading: C:\Documents and Settings\adminstrator\Desktop\Baseword123\BASEWRD3.txt

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
One	258/84.59	132/82.50	109
Two	31/10.16	22/13.75	21
Three	0/ 0.00	0/ 0.00	0
Not in the lists	16/ 5.25	6/ 3.75	?????
Total	305	160	130

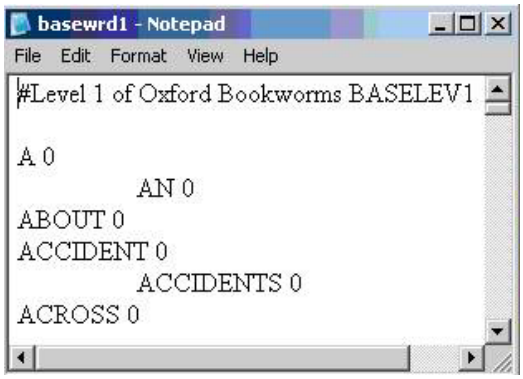
A marked text can show which baseword lists each word in the input text belongs to. It can be used to examine all the words in the text in detail. For example, in the following, unmarked words were in the first list, words marked with <2> were in the second list, words marked with <3> were in the third list, and words marked with <!> were not in any of the lists. This helps check the words that might affect reading ability at a certain level.

> ONE SATURDAY {2}AFTERNOON IN A SMALL TOWN, {!}EMMA {!}CARTER CAME OUT OF A {2}SHOE {2}SHOP WITH SOME NEW {2}SHOES. THEY WERE {2}CHEAP {2}SHOES, BUT {!}EMMA WAS VERY PLEASED WITH THEM. SHE WAS SEVENTY THREE YEARS OLD AND DID NOT HAVE MUCH MONEY. SHE BEGAN TO WALK HOME. 'A {2}NICE {2}CUP OF {2}TEA,' SHE THOUGHT, 'AND THEN I CAN GO FOR A WALK IN MY NEW {2}SHOES.'
>
> IT WAS A {2}QUIET TOWN AND THERE WAS NOBODY IN THE STREET. {2}SUDDENLY, {!}EMMA HEARD SOMETHING BEHIND HER. SHE DID NOT HAVE TIME TO LOOK, BECAUSE JUST THEN SOMEBODY RAN UP BEHIND HER, {2}HIT HER ON THE HEAD, AND {!}SNATCHED HER {2}BAG OUT OF HER HANDS. {!}EMMA FELL DOWN ON HER BACK. THEN SHE LOOKED UP, AND SAW A {2}TALL YOUNG MAN WITH LONG, {2}DIRTY {2}BROWN {2}HAIR. HE STOOD AND LOOKED DOWN AT HER FOR A SECOND; THEN HE RAN AWAY WITH {!}EMMA'S {2}BAG UNDER HIS ARM.
>
> 'HELP! HELP!' {!}EMMA CRIED. BUT NOBODY CAME, AND AFTER TWO OR THREE MINUTES {!}EMMA {2}SLOWLY GOT UP AND WENT TO THE NEAREST HOUSE.
>
> THE PEOPLE THERE WERE VERY KIND. THEY GAVE {!}EMMA A {2}CUP OF {2}TEA, AND SOON AN {!}AMBULANCE CAME AND TOOK HER TO {2}HOSPITAL. AT THE {2}HOSPITAL A DOCTOR LOOKED AT {!}EMMA'S HEAD AND BACK. 'YOU'RE GOING TO BE {!}OK,' HE SAID. 'JUST TAKE IT EASY FOR A DAY OR TWO. CAN YOUR HUSBAND HELP YOU AT HOME?'
>
> 'MY HUSBAND DIED EIGHT YEARS AGO,' SAID {!}EMMA. 'THERE'S ONLY ME AT HOME.' 'WELL,' THE DOCTOR SAID, 'WE DON'T WANT YOU TO FEEL ILL AND FALL {2}DOWNSTAIRS AT HOME. SO I THINK YOU MUST STAY IN {2}HOSPITAL FOR {2}TONIGHT, AND PERHAPS {2}TOMORROW NIGHT, TOO.' LATER, A {2} POLICEMAN CAME TO THE {2}HOSPITAL AND {!}EMMA TOLD HIM ABOUT THE {2}BAG {!}SNATCHER. 'DID ANYBODY SEE THIS YOUNG MAN?' HE ASKED.

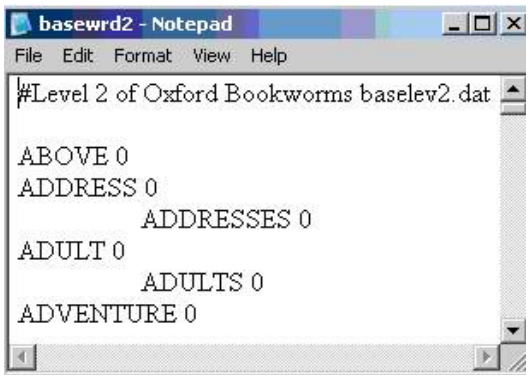
Appendix C

The Six Baseword Lists Established From the Wordlists of the OBW and CER Series

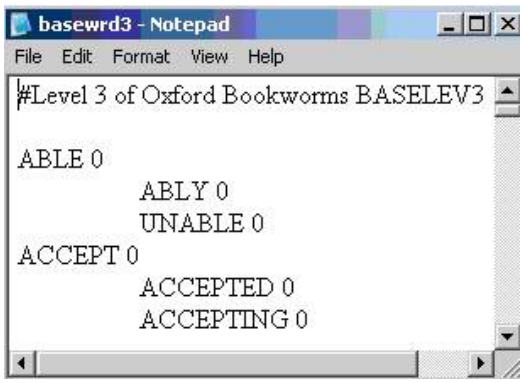
OBW baseword lists. The six OBW baseword lists included all the words at the levels established by the publisher.



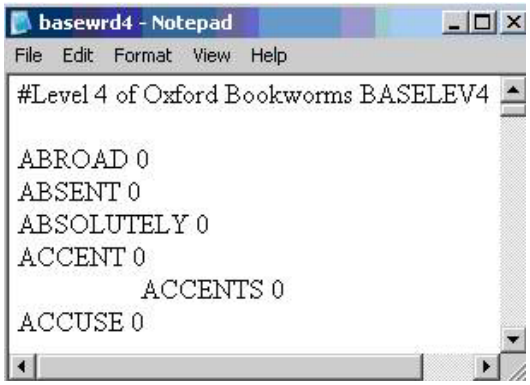
Level 1 OBW basewords



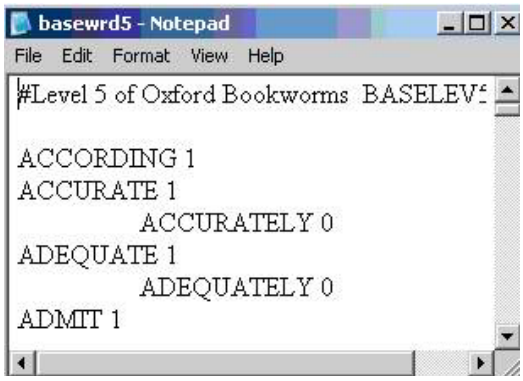
Level 2 OBW basewords



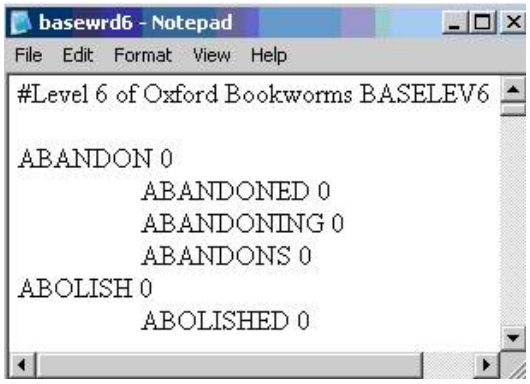
Level 3 OBW basewords



Level 4 OBW basewords

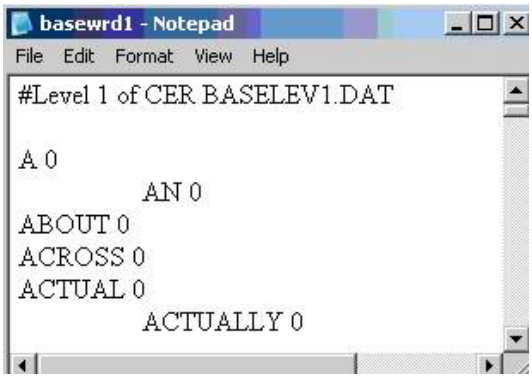


Level 5 OBW basewords

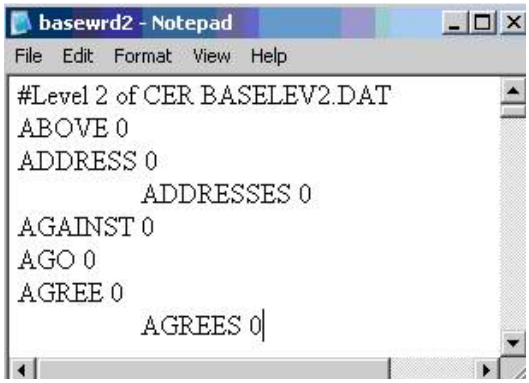


Level 6 OBW basewords

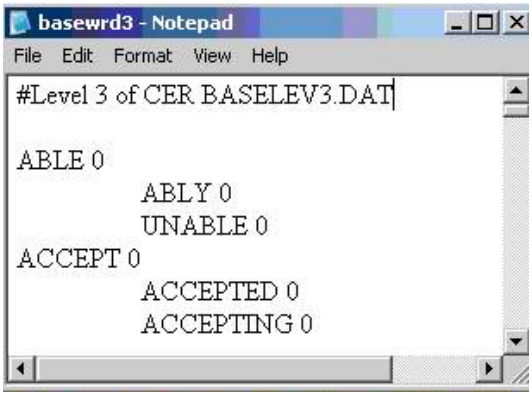
CER baseword lists. The six CER baseword lists included all the words at the levels established by the publisher.



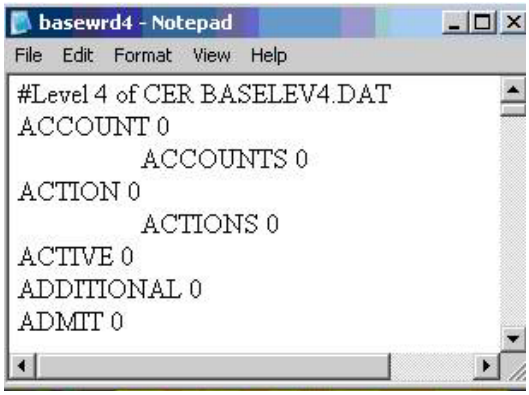
Level 1 CER basewords



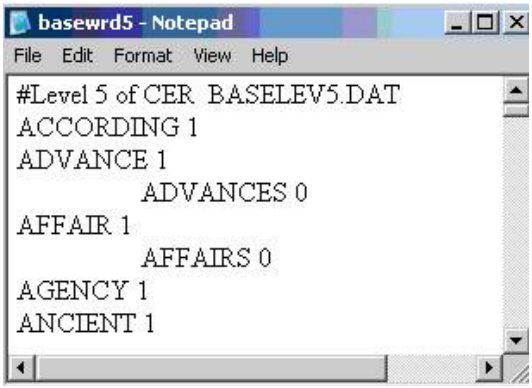
Level 2 CER basewords



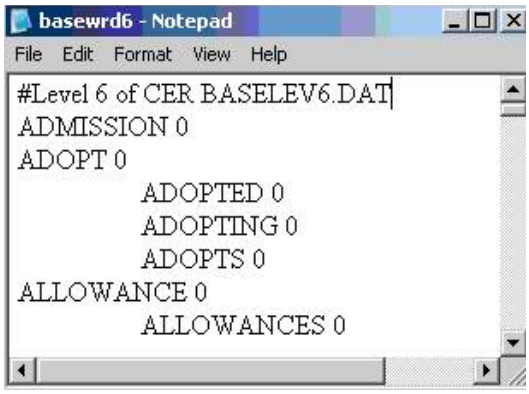
Level 3 CER basewords



Level 4 CER basewords



Level 5 CER basewords



Level 6 CER basewords

About the Author

Udorn Wan-a-rom is an assistant professor and a full-time lecturer at the Department of Western Languages and Linguistics, Faculty of Humanities and Social Sciences, Mahasarakham University, Thailand. He received his MA and PhD in applied linguistics from the Victoria University of Wellington, New Zealand. His main research is in second language (L2) reading, L2 testing, and L2 vocabulary acquisition.
E-mail: romud2505@yahoo.com